
統計學習 (108-2) 期末專案

2020-04-13; Version 1.0

期末專案的目的是希望讓同學透過這個專題，練習並了解統與掌握機器學習與資料分析的各種技巧。除了本文件，請參考 [Hands-On_Machine_Learning]中的第二章以及 **Appendix B**。本說明中的步驟與綱要主要來自前述兩個來源。

一個專案主要包含八個部分 (cf. Appendix B)：

- Frame the problem and look at the big picture.
- Get the data.
- Explore the data to gain insights.
- Prepare the data to better expose the underlying data patterns to Machine Learning algorithms.
- Explore many different models and short-list the best ones.
- Fine-tune your models and combine them into a great solution.
- Present your solution.
- Launch, monitor, and maintain your system.

每個專案都應該要包含前述的七大項的內容。最後一項可以忽略。與 **Appendix B** 以及第二章重複的部分不在此贅述。

分組

自行分組，每組三至六人，請取一個組名。分組資訊請寄給助教：李羽浩 (r07725034@ntu.edu.tw) 或陳思鈺 (r08725043@ntu.edu.tw)。如果你在期限內沒有分組，將會被隨機分組。

分析主題與建構資料集

在進行專案前，需決定分析的主題與所需的資料集是否可以取得。這兩件事可以一起決定，或先後決定。使用的資料集必須要能回答你想要分析的問題。下面列出幾個常用的資料集來源

- UCI Machine Learning Dataset: <http://archive.ics.uci.edu/ml/index.php>
- Kaggle: <https://www.kaggle.com/>
- Taiwan Government Open Data: <https://data.gov.tw/>

如果你使用現成的資料集，需註明資料的來源（如提供者與網址）。如果你分析的問題與方法是參考其他人的專案，需註明來源，並說明你的過程、結果與參考專案的異同。沒有適當說明者將會被視為抄襲並扣分。如果一個專案與其他專案(不論是網路上找到的或是其他同學的) 100%相同或極為相似，這個專案將會失去大部分的分數。

計劃書

各組須在指定期限內繳交計劃書。計劃書的長度為 1–2 頁 A4 紙。內容須包含：

- 組員姓名與學號。
- 分析主題 (How you frame the problem? What is the big picture?)。
- 資料集來源、簡單描述 (以文字描述資料集的特性)、資料筆數、欄位數量、欄位名稱與說明、5 至 10 筆的範例資料。

計劃書請在時限內上傳至作業區"計劃書(第一版)"位置，並至下面位置登記

2020-4-28 上午 9:10 至中午 12:20 各組與老師討論計畫書內容的時段:

https://docs.google.com/spreadsheets/d/1OCTiaPfapwYc0T169u3CZqs_h7H1Yi5xmNbi3wmiZzQ/edit?usp=sharing

錄製與上傳報告內容

報告的內容應包含前述機器學習專案的七大項目。每組報告的時間為 20 - 25 分鐘。由於時間有限，報告時需掌握重點。請自行使用螢幕錄製程式錄製影片，輸出為 mp4 格式，並上傳至 COOL 作業區"期末報告影片上傳(mp4)"。

錄製程式可以使用 Evercam (至 my.ntu.edu.tw 下載)或其他螢幕錄影程式如 Sharex (<https://getsharex.com/>)或 Camtasia (<https://www.techsmith.com/video-editor.html>)。Camtasia 須付費但提供一個月試用。Camtasia 有提供剪接功能。另也可使用 Openshot (<https://www.openshot.org/>)剪接。

詳細的專案內容請整理在 Jupyter Lab 中。為了方便同學閱覽，請將 IPYNB Export 成 HTML 格式後上傳至 COOL"期末報告 HTML 檔案上傳"。另外 IPYNB 檔案、資料檔、其他相關文件請使用 7z 打包後上傳至 COOL 作業區 "期末報告檔案上傳"。

規定

- 未在時限內繳交者將依遲交程度打折。遲交一個小時內期末報告成績 $\times 0.9$ ，遲交兩個小時內期末報告成績 $\times 0.8$ ，依此類推。
- 期末報告的成績中的 30%由為組間互評決定，剩餘 70%由授課教師決定。評分執行細節另行公告。

評分標準

- 分析主題的新穎性與有趣的程度。
- 資料集建構的難度。
- 分析的正確性與完整性。
- 報告流暢度、精確表達、投影片美感等。

重要日期

- 2020-4-20 下午 11:59 以前: Email 分組給助教最後期限
- 2020-4-21 下午 11:59 以前: 助教公布分組名單
- 2020-4-27 下午 11:59 以前: 繳交計劃書至 COOL 作業區，並登記 4/28 計劃書實體討論時段。
- 2020-4-28 上午 9:10 至中午 12:20: 各組與老師討論計畫書內容。

- 2020-5-5 上午 9:10 以前: 繳交第二版計劃書 (指定組別) 至 COOL 作業區。
- 2020-6-16 中午 12:20 以前: 上傳報告影片、IPYNB Export HTML，IPYNB 檔案與資料檔至 COOL 作業區，並線上登記組內互評分數。
- 2020-6-17 中午 12:20 以前: 開放期末報告影片線上觀看。
- 2020-6-20 中午 12:20 以前: 線上登記所有修課同學對各組期末報告的評分。
- 2020-6-23 中午 12:20 以前: 公布各組期末報告分數。
- 2020-6-25 中午 12:20 以前: 接受期末專案評分查詢與更正。本時點之後期末專案成績將不會再更動。